Universidad Autónoma de Madrid
Escuela politécnica superior



Máster en Bioinformática y Biología Computacional

Trabajo Fin de Máster

# TFEA.CHIP: UNA HERRAMIENTA PARA ANALIZAR EL ENRIQUECIMIENTO DE FACTORES DE TRANSCRIPCIÓN APROVECHANDO DATOS DE CHIP-SEQ

Autor: Laura Puente Santamaría
Tutor: Luis del Peso Ovalle

Febrero 2019

# TFEA.CHIP: UNA HERRAMIENTA PARA ANALIZAR EL ENRIQUECIMIENTO DE FACTORES DE TRANSCRIPCIÓN APROVECHANDO DATOS DE CHIP-SEQ

Autor: Laura Puente Santamaría
Tutor: Luis del Peso Ovalle

Laboratorio 252 IIBM
Dpto. de Bioquímica
Facultad de Medicina
Universidad Autónoma de Madrid
Febrero 2019

# Resumen


## Resumen

La identificación de factores de transcripción (FT) responsables de la co-regulación de un conjunto de genes es un problema común en transcriptómica. Con el desarrollo de TFEA.ChIP se busca ofrecer una herramienta para estimar y visualizar el enriquecimiento de FT en un grupo de genes diferencialmente expresados que tenga en cuenta las variaciones en el comportamiento de FTs entre diferentes tipos celulares y estímulos. Con ese fin, se han reunido experimentos de ChIP-seq del consorcio ENCODE y el repositiorio GEO Datasets, y, combinándolos con información de sitios sensibles a DNasa y regiones *enhancer*, se ha generado una base de datos relacionando FTs con los genes con los que regula en cada experimento de ChIP-seq.

En su estado actual, TFEA.ChIP incluye 1154 ChiP-Seqs en células humanas, abarcando 333 FTs diferentes. TFEA.ChIP acepta como entrata (*input*) tanto grupos de genes diferencialmente expresados como listas que incluyan todo el transcriptoma ordenado por la magnitud de la variación en la expresión entre las condiciones comparadas. A partir de esta entrada calcula una puntuación de enriquecimiento para cada uno de los experimentos almacenados en la base de datos interna. La validación de TFEA.ChIP usando una amplia variedad de conjuntos de genes que representan firmas moleculares revisadas para distintos estados y procesos biológicos, indica que el programa identifica los FTs relevantes entre los 10 primeros candidatos en 38 de 49 ocasiones, y entre los 5 % mejores en 45 de 49, alcanzando como predictor un área bajo la curva de 0,89. Además, mediante el análisis detallado de conjuntos de datos de RNA-seq, se ilustra que el uso de métodos basados en datos de ChIP-seq en vez de matrices de peso posicionales permite expandir el análisis de enriquecimiento de FT para incluír modificadores de cromatina y co-factores que carecen de dominio de unión a ADN, además de proporcionar un contexto biológico para inferir el comportamiento de FT dependiente de las condiciones de estímulo o del tejido.

Para facilitar su integración en protocolos de análisis transcriptómicos, así como permitir la expansión y personalización de la base de datos relacionando FTs con genes de forma sencilla, se ha implementado TFEA.ChIP como paquete de R. Además, para hacer la herramienta accesible a un mayor número de investigadores, también se ha desarrollado una aplicación web que ejecuta el paquete desde el servidor, permitiendo así realizar análisis exploratorios de forma sencilla a través de una interfaz gráfica.

TFEA.ChIP está disponible en Bioconductor, GitHub y como aplicación web. Está disponible también una versión preliminar del artículo describiendo TFEA.ChIP en bioRxiv.



## Palabras Clave

Bioinformática, Transcripción, Genómica, Regulación Expresión Génica, Factor de Transcripción, ChIP-Seq, RNA-Seq, Enriquecimiento de factores de transcripción.

# Abstract

The identification of transcription factors (TF) responsible for the co-regulation of an specific set of genes is a common problem in transcriptomics. With the development of TFEA.ChIP we aim to provide a tool to estimate and visualize TF enrichment in a set of differentially expressed genes that takes into account the wide variation in TFâĂŹs behavior across different cell types and stimuli. To that end, we gathered ChIP-Seq experiments from the ENCODE Consortium and GEO Datasets, and used data from Dnase Hypersensitive Sites across cell lines and enhancer location to generate a database linking TFs with the genes they regulate in each ChIP-Seq experiment.

In its current state, TFEA.ChIP includes 1154 ChIP-seq experiments in human cells, covering 333 transcription factors. TFEA.ChIP takes as input differentially expressed gene sets as well as lists including the whole transcriptome sorted by its expression change between conditions. Using this input TFEA.ChIP computes an enrichment score for each of the datasets in its internal database. TFEA.ChIP's validation process, using a wide range of gene sets representing curated molecular signatures of different biological states and processes, indicates that the software identifies relevant TFs within the top 10 candidates in 38 out of 49 tested sets, and within the top 5 % candidates in 45 out of 49, reaching an area under the curve of 0.89 as a predictor. In depth analysis of RNAseq datasets illustrates that the use of ChIP-Seq based methods instead of position weight matrices allows to expand the analysis of TF enrichment to include chromatin modifiers and co-factors that lack a DNA binding domain, in addition to provide a biological context to infer tissue and stimuli-dependent TF behavior.

To facilitate its integration into transcriptome analysis pipelines and allow easy expansion and customization of the TF-gene database, we implemented TFEA.ChIP as an R package. In addition, to make it available to a wide range of researches, we have also developed a web application that runs the package from the server side and enables easy exploratory analysis through a graphic interface.

TFEA.ChIP is available at Bioconductor, GitHub, and as a web application. A preprint version of the article describing TFEA.ChIP is also available at bioRxiv.



# Key words

Bioinformatics, Transcription, Genomics, Gene Expression Regulation, ChIP-Seq, RNA-Seq, Transcription Factor Enrichment.

# Agradecimientos

A Luis del Peso, gracias por darme la oportunidad de formar parte de este equipo donde he aprendido tanto.

A Benilde Jiménez, Bárbara Acosta, Rosana Hernández y María Tiana, gracias por vuestra labor de mentoría durante estos dos años.

# Índice general

# Índice de Figuras

# 1

# Introducción

En el escenario más sencillo, la comparación del transcriptoma de células u organismos en dos condiciones lleva a la identificación de un conjunto de genes diferencialmente expresados (DEGs, del inglés *Differentially Expressed Genes*). El supuesto subyacente es que uno -o unos pocos- factores de transcripción (FTs) regulan la expresión de estos genes. La cuestión que se aborda en este TFM es la identificación computacional de esos FT responsables de la co-regulación de de los DEG.

Hasta hace poco tiempo la aproximación principal para predecir sitios de unión de FT (TFBS, del inglés *Transcription Facto Binding Sites*) situados cerca de los DEGs ha sido utilizar matrices de peso posicionales (*PWMs*, del inglés *Position Weight Matrices*). Estas estructuras representan el sitio de unión de un FT utilizando una matriz en la que una de las dimensiones representa los símbolos del alfabeto (en el caso de DNA: A, C,G y T) y la segunda dimensión las posiciones en la secuencia reconocida por el FT, siendo cada entrada de la matriz la frecuencia con la que se observa cada símbolo en una posición determinada (ver "PWM.[en] la figura 1.1). La comparación de una secuencia dada con una PWM permite calcular una puntuación de similitud simplemente sumando los valores de la matriz correspondientes a cada uno de los símbolos de la secuencia para cada posición. Si la puntuación supera un umbral, se concluye que la secuencia dada contiene un sitio de unión para el FT representado por la PWM. La aplicación de este método a un conjunto de secuencias correspondientes a DEGs frente a secuencias de genes que no responden al estímulo testado[1], permite determinar si la PWM (y por tanto el FT) está sobrerrepresentada en el conjunto de genes DE (ver figura 1.1).

Este método está limitado por varios factores: por un lado, la longitud de los TFBS frente al genoma, así como su carácter degenerado (en cada posición se aceptan varios nucleótidos), da lugar a una alta frecuencia de falsos positivos; por otro, sólo es aplicable a factores de transcripción que se unen directamente a secuencias específicas de ADN y para los que el sitio de unión está caracterizado, excluyendo diversos co-reguladores.

Para paliar estas limitaciones, actualmente comienzan a surgir métodos que aprovechan la gran cantidad de información que se está acumulando en relación a la localización de los sitios de unión de factores de transcripción al genoma determinados de forma experimental [2][3][4][5] mediante Inmunoprecipitación de cromatina seguida de secuenciación (ChIP-seq). Éste es un método experimental desarrollado para analizar interacciones proteína-DNA, y utilizado para cartografiar los sitios de unión de una determinada proteína a lo largo del genoma. Sin embargo, el uso de esta información para predecir los FT responsable de la co-regulación de genes es todavía

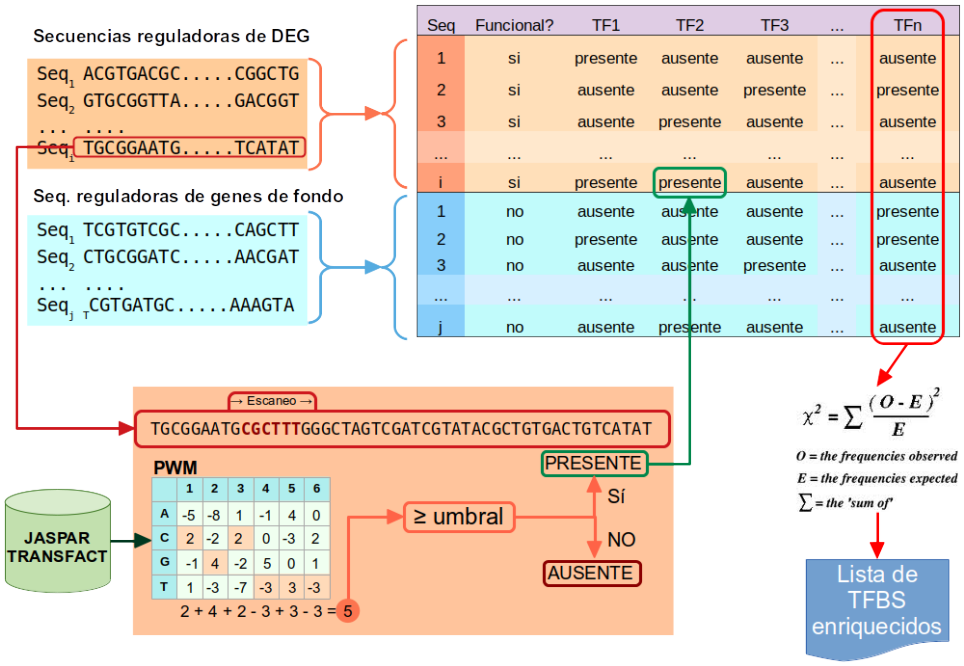


Figura 1.1: **Análisis de enriquecimiento de TFBS mediante el uso de PWMs.** En esta ilustración se muestra el proceso de estimación del enriquecimiento de TFBS en un conjunto de secuencias de DEGs frente a un conjunto de fondo. Teniendo la PWM para un determinado FT (por ejemplo, aquellas almacenadas en bases como JASPAR o TRANSFAC), se estima si hay algún sitio de unión presente en cada una de las secuencias. Por último, se compara el número de TFBS encontrados en cada grupo de secuencias, comprobando si hay sobrerrepresentación en el grupo de interés.

un área en desarrollo que debe resolver varios desafíos para que se generalice como herramienta de estándar de análisis. Una de las principales limitaciones es que aún no se dispone de un método de referencia para asignar dianas génicas a un FT partir de datos de unión a DNA, es decir, conocidos los sitios de unión de un FT a lo largo del genóma es difícil saber con qué gen o genes concretos interacciona (y por tanto regula) cada sitio ya que, en muchos casos, el gen regulado está muy distante del sitio de unión, incluso saltando genes más cercanos. Otro problema a resolver es que algunas de las estrategias que se han comenzado a usar requieren utilizar como entrada datos en formatos muy específicos, en lugar de un formato genérico. Por último, ninguno de ellos está implementado en R, que es el software de referencia para análisis genómicos.

En esta situación, se plantea desarrollar una aplicación bioinformática implementada en R que, aprovechando la información disponible en repositorios públicos[6][7][8], ayude a estimar qué factores de transcripción pueden ser responsables de los cambios en expresión génica a nivel transcripcional .

# 2

# Diseño e implementación

TFEA.ChIP incluye herramientas de análisis y visualización enfocados a la identificación de TFBS enriquecidos en un conjunto de genes. Con este fin, el paquete utiliza información derivada de 685 experimentos de ChIP-Seq generados por el consorcio ENCODE[6], además de 469 ChIP-Seqs realizados por investigadores independientes depositados en GEO[7][8], testando un total de 333 FTs diferentes en una amplia variedad de condiciones experimentales. La colección de FT incluye 278 factores específicos de secuencia y 51 co-reguladores que se unen al ADN de forma indirecta (p. ej. co-factores de transcripción y modificadores de cromatina) o independientemente de la secuencia (p. ej. polimerasas). Así, esta base de datos comprende al rededor del 20 % de los 1391-1600 factores de transcripción específicos de secuencia [9][10] codificados en el genoma humano.

## 2.1. Construcción de la base de datos

Los experimentos de ChIP-Seq publicados contienen las coordenadas cromosómicas a las que se ha unido el FT de interés a lo largo del genoma. De esta forma, para utilizar esta información es necesario asociar las regiones de unión -picos de ChIP-seq- a loci de genes específicos.

En ausencia de información sobre contactos tridimensionales, como la que se produce en experimentos de captura de la conformación de cromatina (Hi-C), lo habitual es asignar cada pico al gen más cercano. Sin embargo, diversas evidencias experimentales indican que sólo una pequeña fracción de las interacciones entre regiones distantes son con el gen más próximo[11]. De acuerdo con esto, la incertidumbre en la asignación de picos a genes aumenta conforme mayor es la distancia entre el sitio de unión y su gen diana[12].

Para superar estos obstáculos hemos seguido una estrategia en tres pasos: en primer lugar definimos todos los sitios con potencial regulador de la expresión génica a nivel genómico (RPRT, Región con Potencial Regulador de la Transcripción); a continuación, asignamos cada una de estas RPRTs a genes concretos, estableciendo pares RPRT-gen; por último, asignamos sitios de unión de ChIP-seq a genes de acuerdo a su co-localización con alguno de los pares RPRT-gen. La accesibilidad de la cromatina es requisito obvio para que una región sea accesible a la maquinaria transcripcional y, por tanto, tenga capacidad de regular la expresión génica. Puesto que la cromatina accesible, presenta sensibilidad a DNasa, los sitios sensibles a DnasaI (DHS) se

consideran tradicionalmente como un marcador de región reguladora. Así, para el primer paso definimos todos los DHS recogidos en el proyecto ENCODE[13][14] localizados a menos de 1kb del gen, como RPRT proximales (Figura 2.1 A). Por otra parte, definimos las RPRT distales como las DHS que solapan con los enhancer almacenados en la base de datos GeneHancer (v. 4.7)[15] (Figura 2.1 B).

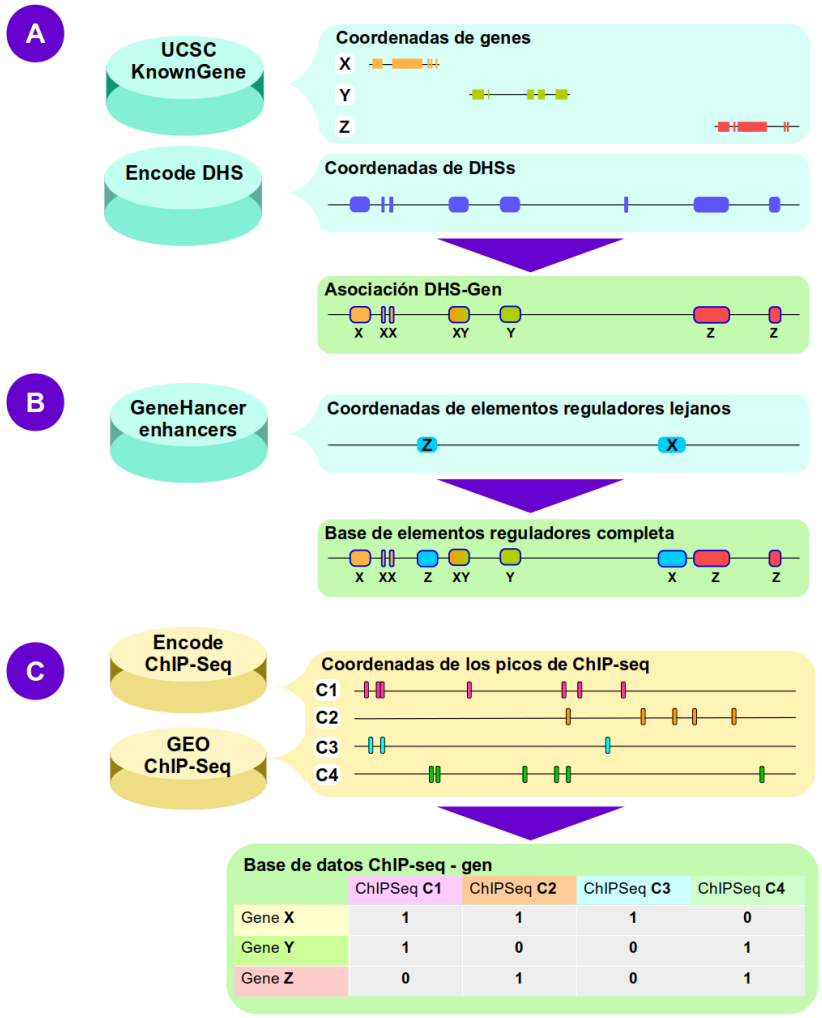


Figura 2.1: **Construcción de una base de datos asociando FTs a conjuntos de genes. A:** Se seleccionan regiones abiertas de cromatina (definidas como clústers de DHS identificadas en el proyecto ENCODE) que se sitúen a 1Kb o menos de alguno de los genes en KnownGene. **B:** se añaden los elementos reguladores distantes asociados a genes recogidos en GeneHancer, completando la base de regiones reguladoras asociadas a genes. **C:** se seleccionan los picos de cada ChIP-Seq que solapan con alguna de las regiones reguladoras, asignando los genes correspondientes a cada ChIP-Seq.

En el segundo paso, cada RPRT proximal se asignó a los genes solapantes o adyacentes hasta 1Kb de distancia, permitiendo que una misma RPRT proximal sea asociada a dos genes si solapa con ambos. Este proceso resulta en una colección de pares RPRT proximal-gen, manteniendo sólo aquellas RPRT relacionados con al menos un gen (Figura 2.1A). Para asignar las RPRT distales, utilizamos la información de GeneHancer[15]. Esta base de datos integra diversas evidencias experimentales y conocimiento experto, para definir regiones del genoma como potenciales enhancer y asociarlos a genes específicos. Partiendo de esta información, seleccionamos las regiones descritas en GeneHancer con una alta probabilidad de función enhancer (GH

score >1) y elevado coeficiente de correlación con algún gen diana (Gene Association Score > 10)(Figura 2.1B). Cada RPRT distal solapante con un enhancer se asocia al gen o genes con los que el enhancer correlaciona.

Finalmente, para cada experimento de ChIP-Seq, se seleccionan aquellos picos que sean estadísticamente significativos (FDR $< 0.05$) y solapen con alguna de las regiones reguladoras mencionadas anteriormente (tanto proximales como distales), asignando cada pico al gen que potencialmente regula la RPRT (Figura 2.1C). Por último, se integra la información de todos los ChIP-Seq en una matriz, con filas que se corresponden con los genes recogidos en KnownGene y una columna para cada experimento de ChIP-Seq analizado; asignando un 1 para los experimentos que tengan al menos un pico asignado a ese gen, o un 0 si no lo tienen.

## 2.2. Análisis del enriquecimiento

TFEA.ChIP se ha diseñado para tomar el resultado de un análisis de expresión diferencial e identificar FTs enriquecidos en la lista de DEGs. Se parte de la premisa de que los efectores clave de una respuesta transcripcional tendrán más dianas entre los DEGs que entre un grupo de genes que no responda al estímulo de interés. Para realizar esta tarea, TFEA.ChIP implementa dos tipos de tests para identificar FTs enriquecidos: análisis de asociación y GSEA.

### 2.2.1. Análisis de asociación

Es el método más sencillo para estimar enriquecimiento y sólo requiere como *input* una lista de genes de interés (en el caso de un análisis de expresión diferencial, los DEGs). Se estima el enriquecimiento a partir de la relación entre los DEGs y los genes diana de un TF determinado en cada ChIP-Seq. Para ello, se construye una tabla de contingencia, de dimensión 2x2, clasificando los genes según su respuesta transcripcional (diferenciando entre DEGs y genes que no responden al estímulo) y la presencia de un sitio de unión para el factor de transcripción (TFBS) asociada a ese gen en el ChIP-seq (Figura 2.2, A.1). La diferencia de proporciones (Odds Ratio, OR) estará sesgada, si el TF es responsable de la respuesta transcripcional. La significancia estadística de la asociación para cada factor de transcripción se determina mediante un test exacto de Fisher. Posteriormente se ajustan los p-valores, para corregir por el testado de múltiples hipótesis, mediante el procedimiento de Benjamini-Hochberg. Finalmente, para ordenar los diferentes experimentos de ChIP-seq en función de su enriquecimiento se define una medida de distancia calculada como:

$$Distancia_i = \sqrt{OR_i^2 + LPV_i^2}, \quad \text{Para valores de } LOR_i > 0$$

$$Distancia_i = \sqrt{\frac{1}{OR_i}^2 + LPV_i^2}, \quad \text{Para valores de } LOR_i < 0$$

Siendo $LOR_i = log_2(\text{Odds Ratio})$ y $LPV_i = log_{10}(\text{p-valor ajustado})$ para el ChIP-seq $i$

Con los datos de distancia TFEA.ChIP genera una tabla con los diferentes experimentos de ChIP-seq almacenados en la base de datos interna, ordenados en base al parámetro de distancia, siendo el experimento de mayor distancia positiva el que muestra un mayor enriquecimiento. Además, el programa representa los resultados en un gráfico de puntos (LOR vs LPV) interactivo que permite la identificación de experimentos individuales (Figura 2.2, A.2).

### Integración de los resultados por FT

Siguiendo el procedimiento anteriormente descrito obtenemos una estimación de enriquecimiento para cada experimento $ChIP - seq_{\mathrm{i}}$, o lo que es lo mismo, para cada $\mathrm{FT_j}$, en un tipo celular y unas condiciones experimentales concretas. Puesto que para muchos FT existen varios experimentos de ChIP-seq individuales, para estimar el comportamiento global del factor frente al estímulo que se está testando es necesario integrar los resultados de todos los experimentos de ChIP-seq correspondientes a cada FT. Para comprobar si hay diferencias significativas en el ranking alcanzado por los ChIP-Seqs asociados al $\mathrm{FT_j}$ frente a los demás, TFEA.ChIP da dos alternativas, por un lado, realizar una prueba de los rangos con signo de Wilcoxon, y por otro, un test de tipo Kolmogorov-Smirnov, similar al utilizado en GSEA[16],[17].

Es importante tener en cuenta que integrar los resultados por factor de transcripción puede apantallar el enriquecimiento estimado en los casos en los que el comportamiento de un factor de transcripción sea muy variable en los diferentes ChIP-seq almacenados en la base de datos de TFEA.ChIP.

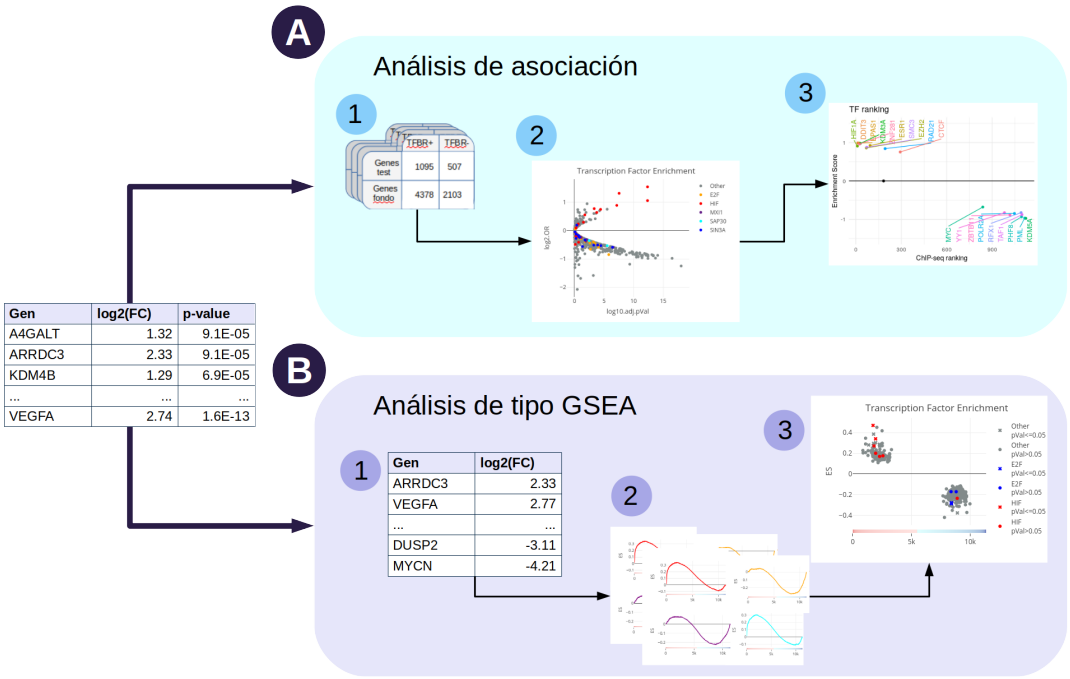


Figura 2.2: **Análisis de enriquecimiento de TFBS.** Partiendo del resultado de un análisis de expresión diferencial (RNA-seq, microarray...) en forma de conjunto de DEG o listado de genes ordenados en base a su diferente expresión en las condiciones analizadas, podemos realizar un análisis de asociación en el primer caso y un análisis tipo GSEA en el segundo. **A:** Análisis de asociación. Partiendo de un conjunto de DEG se genera una matriz de contingencia para cada experimento (A.1), a partir de las cuales se calculan ORs y p-valores (A.2). Posteriormente se pueden integrar los resultados por factor de transcripción (A.3). **B:** Análisis de tipo GSEA. comenzamos ordenando los genes en función de su cambio en la expresión (B.1). Los genes se ordenan en función de su cambio en la expresión (B.1). A partir de esta tabla ordenada, se calcula una puntuación de enriquecimiento para cada experimento ChIP-Seq en la base (B.2, B.3), así como el p-valor para esa puntuación.

### 2.2.2. Análisis de tipo GSEA

El enriquecimiento de los diferentes FTs en un experimento de expresión diferencial puede estimarse también mediante el método utilizado en GSEA (del inglés *Gene Set Enrichment Analysis*). Para implementar este método se utiliza la lista de genes asignada a cada experimento de ChIP-Seq en la base interna como un *gene set*, representando la firma de unión de cada FT bajo unas condiciones experimentales determinadas. Así, cada columna en la figura 2.1C se representa como un *gene set* que incluye todos los genes con un valor de 1. Este análisis requiere como datos de entrada una lista ordenada de genes, así como la variable utilizada para ordenarlos, bien sea *Fold Change*,$log_2$*(Fold Change)* o una medida semejante, como el $\pi$-value[18], que combina en una sola variable el $log_2$*(Fold Change)* y el p-valor de cada gen.

Tal como se describe en [16], dada una lista de genes ordenada por su cambio en expresión $\boldsymbol{G}$ (todos los genes medidos en el experimento de expresión diferencial) y un conjunto de dianas $\boldsymbol{D}$ (el *gene set*), el objetivo es comprobar si aquellos elementos de $\boldsymbol{D}$ se encuentran dispersos de forma aleatoria a lo largo de $\boldsymbol{G}$ o se encuentran desplazados hacia el principio o el final de la lista, entre los genes con mayor variación en los niveles de expresión.

La puntuación de enriquecimiento (ES) se calcula recorriendo la lista $\boldsymbol{G}$, sumando una cantidad por cada elemento que pertenece a $\boldsymbol{D}$, restando en caso contrario. Las cantidades a sumar y restar se determinan en función del número de elementos de $\boldsymbol{D}$ presentes en $\boldsymbol{G}$, así como del valor de la variable utilizada para ordenar los genes, de forma que la ES quede acotada entre (-1,1). De esta forma, los genes presentes en los extremos de $\boldsymbol{G}$ tendrán un mayor peso sobre la puntuación final.

Para cada experimento de ChIP-Seq almacenado en la base interna, TFEA.ChIP devuelve la máxima puntuación de enriquecimiento (ES) y el *Argumento*, esto es, el punto de la lista ordenada de genes en el que se alcanza la ES. Para estimar la robustez de la ES calculada se realiza un test de permutación.

## 2.3. Aplicación web

Para ampliar la accesibilidad de TFEA.ChIP se ha desarrollado una aplicación web en la que el usuario puede acceder a la mayor parte de las funciones que ofrece el paquete de R de forma sencilla a través de una interfaz gráfica interactiva. La aplicación se ha desarrollado utilizando el paquete de R shiny[19].

# 3

# Validación y rendimiento

## 3.1. MSigDB como referencia

Como primera aproximación para validar el funcionamiento de TFEA.ChIP se escogió el repositorio de firmas moleculares Hallmark, parte de la base de datos MSigDB[20]. Este repositorio contiene 50 conjuntos de genes (*gene sets*) curados representando genes clave en distintos procesos y estados biológicos. Tras identificar en la literatura los FTs relacionados con cada uno de los procesos representados por cada uno de los *gene sets* (FT-esperados), se procedió a estimar el enriquecimiento de FTs con TFEA.ChIP.

Para cada *gene set* en Hallmark se realizaron 100 análisis por asociación, utilizando como conjunto control una muestra al azar de 1000 genes del resto del genoma. A partir de los resultados de las 100 iteraciones se calculó la distancia media estimada para cada experimento, evaluando en qué posiciones del ranking quedaban los experimentos de ChIP-Seq de FTs relevantes para el proceso biológico representado en cada *gene set*. Como se muestra en la figura 3.1A, la predicción de TFEA.ChIP incluía al FT-esperado entre los 10 primeros puestos en 38 de los 49 *gene sets* analizados (uno de los *gene sets* no se analizó pues no existe información de ChIP-seq para el FT relevante en ese caso). Si en lugar de considerar los 10 mejores puestos, se consideran clasificaciones correctas aquellas en las que el FT-esperado está entre el 5 % superior de las predicciones, TFEA.ChIP identificó correctamente el FT relevante en 45 de los 49 casos estudiados (Figura 3.1A).

Para estimar la eficiencia de TFEA.ChIP se repitió el proceso construyendo *gene sets* muestreando aleatoriamente todos aquellos genes contenidos en alguno de los conjuntos de Hallmark. De esta forma se calculó la sensibilidad y la especificidad de TFEA.ChIP, alcanzando un área bajo la curva de 0.89, lo que indica que TFEA.ChIP tiene un buen comportamiento como clasificador, esto es, identificando correctamente el FT relevante a partir de un conjunto de DEGs.

Como complemento, se siguió el mismo procedimiento con un conjunto de 124 *gene sets* procedentes del apartado C2 de la base de datos MSigDB, todos relacionados con perturbaciones químicas o genéticas extraídos de literatura biomédica. C2 tiene un proceso de curación menor, por lo que cuenta con mayor variación en la cantidad de genes que incluyen y las condiciones experimentales utilizadas. Como se puede ver en la figura 3.1C y D, este caso, para 98 de los 124 *gene sets* se identificó al FT relevante en el 5 % superior de las predicciones (en 64 de ellos, entre los 10 primeros), alcanzando un área bajo la curva de 0.82.

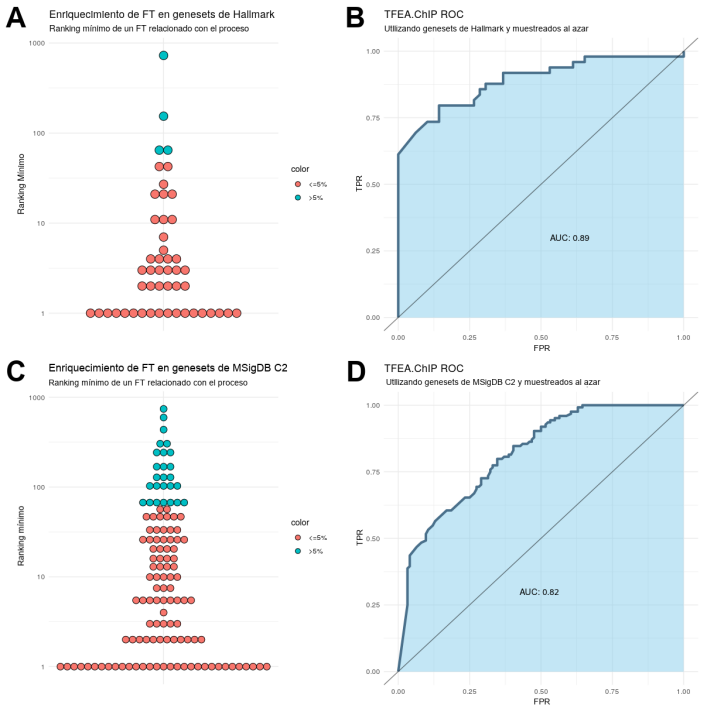


Figura 3.1: **Rendimiento de TFEA.ChIP. A:** Se ejecutó TFEA.ChIP utilizando como entrada cada uno de los 49 *gene sets* valorables y, en cada caso, se anotó la posición que ocupaban los FT-esperados para cada *gene sets* en el listado ordenado de predicciones. Cada uno de los puntos de la gráfica representa el rango (posición) ocupada por el FT-esperado en cada uno de los *gene sets* analizados. Puesto que para un mismo FT la base de datos de TFEA.ChIP puede contener varios experimentos de ChIP-seq, se muestra el la posición del mejor. **B:** Repitiendo el procedimiento con conjuntos de genes escogidos al azar del mismo tamaño se calculó un área bajo la curva de 0,89. En **C** y **D** se muestran los resultados del análisis de 124 *gene sets* procedentes de la base de perturbaciones químicas y genéticas de MSigDB C2

## 3.2. Aplicación de TFEA.ChIP a datos experimentales

Una de las ventajas de TFEA.ChIP es que, en lugar de integrar todos los experimentos realizados con un FT determinado en un listado consenso de dianas, guarda información individualizada de cada experimento de ChIP-seq junto con metadatos de relevancia biológica. Esto permite un análisis detallado del que se puede obtener información relevante para apoyar o rechazar la predicción de un FT. Así, como caso ilustrativo se utilizó un experimento de RNA-Seq depositado en el repositorio GEO con número de *accession* GSE89831, en el que se describe el efecto de la hipoxia en los niveles de mRNA de nueva síntesis. Este análisis además, pondrá de manifiesto el rendimiento de TFEA.ChIP sobre un conjunto de datos experimentales crudos sin ningún tipo de filtrado por parte de expertos como es el caso de los *gene sets* de MSigDB.

La respuesta transcripcional a la falta de oxígeno está mediada por un grupo de factores de tipo hélice-giro-hélice (bHLH) llamados Factores Inducibles por Hipoxia (HIFs). Estos factores HIF forman heterodímeros que comparten una unidad $\beta$ común, codificada por el gen ARNT, y una subunidad $\alpha$, codificada por los genes HIF1A, EPAS1 o HIF3A.

Los datos contenidos en la entrada GSE89931, corresponden al análisis, mediante RNA-seq del efecto de la hipoxia en células endoteliales [21] y, por tanto, esperamos encontrar a los FT de la familia HIF enriquecidos entre los DEGs de este estudio. Así, en primer lugar, descargamos las cuentas por gen (*reads*) e identificamos, con el paquete DESeq2[22], el conjunto de genes cuya

expresión se induce de forma significativa por hipoxia ($log_2(FoldChange) > 0$ y FDR<0.05) y así como un conjunto de genes de control cuya expresión no se alteraba en respuesta a la falta de oxígeno (FDR>0.05). A continuación, ejecutamos TFEA.ChIP utilizando estos conjuntos de genes como entrada. Los resultados que arroja TFEA.ChIP sitúan claramente las tres subunidades HIF por encima del resto de FT disponibles en la base de datos de TFEA.ChIP (Fig 3.2A). De hecho, los experimentos de factores HIF que no muestran enriquecimiento significativo están realizados en condiciones de normoxia, situación en la que los factores HIF están inactivos (Fig 3.2B). Por otro lado, los únicos experimentos normóxicos que muestran enriquecimiento están realizados en la línea de carcinoma renal 786-O. En esta línea el gen supresor de tumores VHL, que favorece la degradación de los factores HIF en presencia de oxígeno, se encuentra mutado, provocando la activación constitutiva de HIF incluso en condiciones normóxicas. Para confirmar que, en su conjunto, los HIF estaban enriquecidos, utilizamos la medida de distancia calculada a partir del OR y el LPV (vease 2.2.1) en un test U de Mann-Whitney, concluyendo que hay un desplazamiento significativo de los ChIPs de factores HIF hacia la parte alta del ranking (p-valor = 4.678e-6).

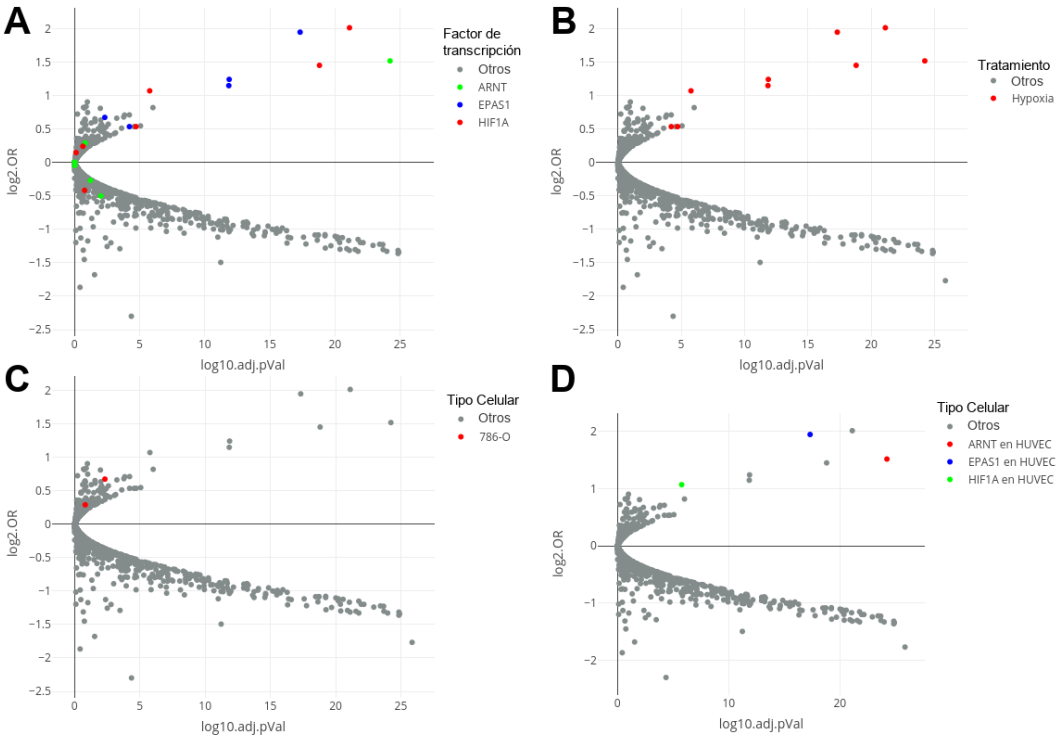


Figura 3.2: **Identificación de los FT responsables de la respuesta transcripcional a hipoxia a partir de datos experimentales. Análisis de asociación**. El conjunto de DEG (genes significativamente inducidos por hipoxia) identificados en los datos de *GSE89831* se analizó con TFEA.ChIP mediante un test de asociación usando como control los genes cuya expresión no cambiaba por hipoxia. Las gráficas muestran el resultado del test asociación representando los valores de LPV frente a LOR para cada uno de los ChIP-seq incluidos en TFEA.ChIP utilizando un código de colores que identifica: **A:** los ChIP-seq realizados con alguna de las subunidad de HIF frente al resto; **B:** los experimentos realizados en hipoxia frente a normoxia; **C:** los experimentos realizados en la línea celular de carcinoma renal 786-O.**D:** Los experimentos de ChIP-seq realizados en HUVEC, la misma línea celular utilizada el experimento de RNA-seq *GSE89831*

Además del análisis de asociación, se realizó un análisis de tipo GSEA con la lista completa de genes del RNA-seq ordenados en base a su respuesta a hipoxia. En la Fig. 3.3A se representa la ES de cada uno de los 1154 *gene sets* de TFEA.ChIP frente al punto de la lista de genes

en el que alcanzan el valor máximo. Los *gene sets* con mayor ES y p-valor ajustado $< 0.05$ se corresponden con aquellos que representan perfiles de unión a DNA de HIFs.

En conjunto, todos estos resultados muestran que los análisis implementados en TFEA.ChIP identificaron con éxito los factores HIF como FT relevantes para la respuesta transcripcional a hipoxia, particularmente en la inducción de genes.

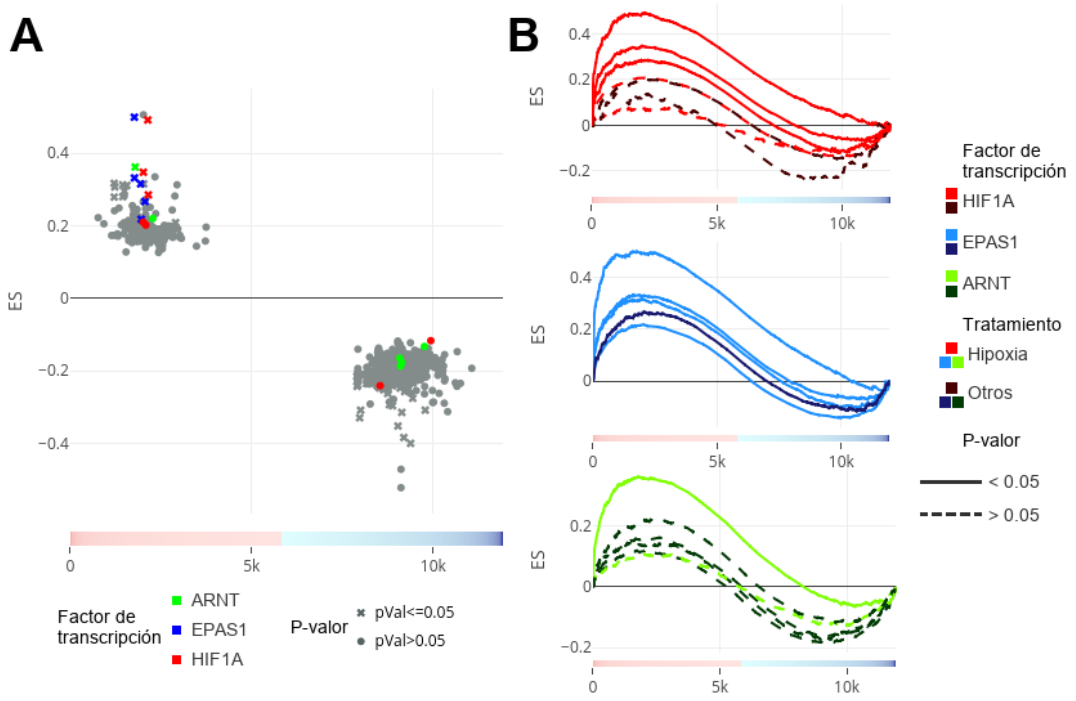


Figura 3.3: **Identificación de los FT responsables de la respuesta transcripcional a hipoxia a partir de datos experimentales. Análisis GSEA.** Los genes expresados en HUVEC, de acuerdo a los datos de *GSE89831*, se ordenaron en base a su respuesta a hipoxia y esta lista se uso como entrada para el análisis de asociación tipo GSEA. **A:** Resultados de ES (*Enrichment Score*) y el argumento (posición máxima de enriquecimiento dentro del listado de genes ordenado) para cada ChIP-seq. Destaca la posición de los factores HIF, significativamente enriquecidos entre los genes más up-regulados por hipoxia. **B:** Puntuación de enriquecimiento de los factores HIF a lo largo de la lista de genes. En todos los casos la máxima puntuación de enriquecimiento es alcanzada por experimentos realizados en hipoxia.

## 3.3. Comparación con otros métodos

Con el objetivo de comprobar si TFEA.ChIP contribuye a superar las limitaciones de los actuales métodos de estimación de enriquecimiento de FT, se comparan los resultados obtenidos a partir de los mismos datos de entrada en otras dos aplicaciones, Opossum[23] y BART[24].

Opossum Single Site Analysis estima el enriquecimiento de FTs buscando TFBS sobrerrepresentados en las secuencias de ADN de un conjunto de genes de interés frente a un grupo de fondo, basándose para ello en el uso de PWMs que describan la secuencia de nucleótidos a la que se une un FT (ver figura 1.1). Este método está inherentemente limitado a FTs que se unen directamente al ADN, pero trabajar con PWMs le da la ventaja de poder conocer todos los potenciales sitios de unión de un FT adyacentes a genes, no sólo aquellos a los que el FT está unido en las condiciones experimentales analizadas en un determinado experimento de ChIP-seq,

siempre que el motivo de secuencia al que se une el FT se haya caracterizado y exista una PWM que lo represente

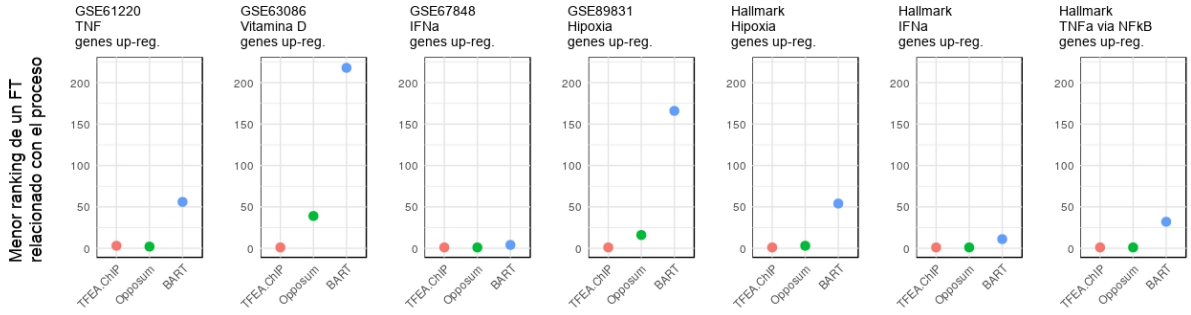


Figura 3.4: **Comparación del rendimiento de TFEA.ChIP con Opossum y BART.** En esta figura se representa el menor ranking obtenido en cada una de las tres aplicaciones para un FT relacionado con el proceso biológico correspondiente. En todos los casos los datos de partida son conjuntos de genes diferencialmente expresados obtenidos de los experimentos de RNA-seq o listas de Hallmark indicados. Como grupo control se utilizó una selección de 1000 genes muestreados al azar para las tres listas de Hallmark, mientras que para los cuatro experimentos de RNA-seq se usan todos aquellos genes que no respondían al estímulo (aquellos con un p-valor $>0.05$). En el caso de Opossum se utilizaron PWMs obtenidas del repositorio *JASPAR CORE*[25] y permitiendo hasta 1Kb de distancia entre los TFBS predichos y los genes. En el caso de MARGE/BART, se siguió el protocolo recomendado de forma que, a partir de cada conjunto de genes, se generó con MARGE una lista de posibles regiones de regulación. Esta lista de zonas reguladoras fue utilizada en BART para estimar el enriquecimiento de FT.

Por otro lado, BART, al igual que TFEA.ChIP, utiliza datos experimentales de ChIP para determinar enriquecimiento de FT. Sin embargo, BART esta diseñado para actuar en conjunto con la aplicación MARGE[26] -en este caso, utilizamos MARGE-cistrome-, un programa que, a partir de experimentos de acetilación de histonas, es capaz de inferir la localización de las regiones reguladoras de genes diferencialmente expresados. Si no se dispone de un ChIP-seq de acetilación de histonas que acompañe al experimento de expresión diferencial, MARGE-cistrome utiliza los datos de experimentos públicos de acetilación. Una vez obtenida la localización de las potenciales regiones reguladoras con MARGE, BART utiliza datos de ChIP-seq de FTs previamente publicados para estimar qué factores de transcripción pueden ser responsables de la respuesta transcripcional observada.

Como puede apreciarse en la figura 3.4, TFEA.ChIP es capaz de igualar o superar el rendimiento de una herramienta establecida como Opossum, añadiendo la posibilidad de estimar el enriquecimiento de factores que se unen de forma inespecífica, como modificadores de histonas (HDACs, KDMs) o que no tienen dominio de unión a DNA (como SIN3A). La diferencia de rendimiento con BART es aún más significativa. En particular, TFEA.ChIP supera a BART en fiabilidad al analizar conjuntos de datos derivados de experimentos de RNA-seq.

En conclusión, TFEA.ChIP tienen un rendimiento superior a estrategias clásicas basadas en PWMs y también mejora las aproximaciones disponibles basadas en el uso de datos de ChIP-seq.

# 4

# Discusión

La identificación de factores de transcripción encargados de coordinar un determinado patrón de expresión es una pieza clave en transcriptómica. El *software* descrito en este trabajo, TFEA.ChIP, combina perfiles de unión FT-ADN determinados experimentalmente con mapas de regiones hipersensibles a Dnasa y regiones enhancer para identificar factores de transcripción enriquecidos.

Una de las características destacadas de TFEA.ChIP es que, en lugar de asignar diréctamente cada pico de ChIP-seq al gen más cercano, incorpora la información sobre enhancers recopilada en GeneHancer, maximizando así la información que se puede obtener del comportamiento de cada FT manteniendo a la vez un nivel razonable de certidumbre al incorporar regiones reguladoras distantes.

La comparación con BART revela una de las particularidades de nuestro trabajo en un campo aún en desarrollo; gran parte de las aplicaciones publicadas, pese a que se basan en el uso de la misma información -experimentos de ChIP-seq- y tratan de estimar un mismo parámetro -enriquecimiento de FT-, se han desarrollado para un tipo de datos de entrada concreto. Por este motivo se da el riesgo de que las aplicaciones sean poco flexibles y no consigan mantener el rendimiento cuando se utiliza un *input* diferente del previsto. Este es uno de los motivos por los que se decidió que los datos de entrada utilizados en TFEA.ChIP fueran listas de IDs genes, ya que, por un lado, hay una amplia gama de protocolos disponibles para relacionar distintos tipos de resultados experimentales (microarrays, RNA-seq, ChIP-seq, etc.) con IDs de genes. Por otro lado, al mantener un formato sencillo y universal para los datos de entrada (IDs de genes en los tres formatos más comunes, junto con la variable numérica utilizada para ordenarlos en caso del análisis tipo GSEA), TFEA.ChIP mantiene una gran flexibilidad, pudiendo utilizarse para identificar enriquecimiento de FTs tanto a partir de resultados de RNA-seq, microarrays, ChIP-seq, como de firmas moleculares, como puede verse en los apartados 3.1 y 3.3.

Otro aspecto diferencial de TFEA.ChIP es que la base de datos utilizada para estimar el enriquecimiento de FT es personalizable. El paquete de software incluye herramientas para que el usuario pueda complementarla con sus propios datos, o incluso generar una base desde cero con su propio criterio para asignar picos de ChIP-seq a genes.

Por último, otras herramientas como *ENCODE ChIP-seq significance tool*[4], iREGULON[5] o ChEA[2] están implementadas como aplicaciones web (en el caso de iREGULON, como *plugin* de Cytoscape, mientras que ChEA está integrado en la suite Enrichr[27]) o son parte de un

protocolo de análisis específico como BART. En cambio, TFEA.ChIP es un paquete ligero de R, que puede ser integrado fácilmente con otras librerías utilizadas en transcriptómica manteniendo un alto grado de flexibilidad y la posibilidad de personalizar la base de datos interna para adaptarse a las necesidades del usuario. Además, para aumentar la disponibilidad del software y hacerlo accesible a una mayor variedad de investigadores, hemos desarrollado una aplicación web interactiva desde la que se puede acceder a la mayoría de las funciones de TFEA.ChIP en un entorno interactivo a través de una interfaz gráfica sencilla.

# 5

# Perspectivas Futuras

Uno de los retos principales de cara al futuro es el mantenimiento de TFEA.ChIP a largo plazo. Con el avance de las técnicas para estudiar la arquitectura tri-dimensional de la cromatina o para descubrir y anotar regiones reguladoras, así como la publicación de nuevos datos experimentales, será necesario revisar el procedimiento seguido en TFEA.ChIP para relacionar experimentos de ChIP-seq y sus dianas. También forma parte de estas tareas el mantenimiento de un catálogo de experimentos de ChIP-seq actualizado y variado, de forma que se incluya una gama representativa y completa de los factores de transcrpción existentes tanto en humano como en ratón, a ser posible en diversos contextos experimentales.

Otro aspecto clave para mejorar TFEA.ChIP es el rendimiento del análisis tipo GSEA, particularmente a la hora de estimar el p-valor de la puntuación de enriquecimiento. Actualmente se lleva a cabo un test de permutación barajando la posición de los genes del *input*, considerando que:

$$\text{p-valor} = \frac{\mathbb{1}(|ES_{perm}| > |ES_{test}|)}{\text{n}^\circ\text{de permutaciones}}$$

Éste es un proceso que consume una cantidad considerable de tiempo, sólo explora una pequeña fracción de las formas en las que podría estar ordenada la lista de genes utilizada, y limita el rango de p-valores resultantes. En la literatura se pueden encontrar algunas publicaciones en los que se plantean soluciones para superar estas limitaciones [28] [29], pero no se ha llegado a definir método analítico para estimar la significancia de las puntuaciones de enriquecimiento. Por estos motivos resultaría beneficioso desarrollar una forma más adecuada de estimar el p-valor de las ESs calculadas para cada experimento en TFEA.ChIP

# 6

# Glosario de acrónimos

- **FT**: Factor de Transcripción
- **DEG**: Differentially Expressed Gene, gen diferencialmente expresado
- **PWM**: Position Weight Matrix, matriz de peso posicional
- **TFBS**: Transcription Factor Binding Site, sitio de unión de factor de transcripción
- **TFBR**: Transcription Factor Binding Region, zona de unión de factor de transcripción
- **DHS**: Dnase Hypersensitive Site, sitio sensible a Dnasa
- **OR**: Odds Ratio, diferencia de proporciones
- **ES**: Enrichment Score, puntuación de enriquecimiento
- **HIF**: Hypoxia Inducible Factor, factor inducible por hipoxia

# Bibliografía


[1] WW Wasserman and A Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5:276, 2004.

[2] Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I. Berger, Amin R. Mazloom, and Avi Ma'ayan. Chea: transcription factor regulation inferred from integrating genome-wide chip-x experiments. *Bioinformatics*, 26:2438–2444, 2010.

[3] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, , and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28:495–501, 2010.

[4] R. K. Auerbach, B. Chen, and A. J. Butte. Relating genes to function: identifying enriched transcription factors using the encode chip-seq significance tool. *Bioinformatics*, 29:1922–1924, 2013.

[5] Rekin's Janky et al. iregulon: From a gene list to a gene regulatory network using large motif and track collections. *PLOS Computational Biology*, 10(7):1–19, 2014.

[6] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2004.

[7] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30:207–210, 2002.

[8] Tanya Barrett et al. NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[9] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10:252–263, 2009.

[10] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. The human transcription factors. *Cell*, 172(4):650 – 665, 2018.

[11] Amartya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489:109–113, 2012.

[12] Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip A Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George A Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nature*, 47:598–606, 2015.

[13] Sam John et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43:264–268, 2011.

[14] Robert E. Thurman et al. The accessible chromatin landscape of the human genome. *Nature*, 489:75–82, 2012.

[15] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, Asher Kohn, Michal Twik, Marilyn Safran, Doron Lancet, and Dana Cohen. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database*, 2017, 2017.

[16] Aravind Subramanian, Pablo Tamayo, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[17] Vamsi K Mootha et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

[18] Yufei Xiao, Tzu-Hung Hsiao, Uthra Suresh, Hung-I Harry Chen, Xiaowu Wu, Steven E. Wolf, and Yidong Chen. A novel significance score for gene selection and ranking. *Bioinformatics*, 30(6):801–807, 2014.

[19] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2018. R package version 1.2.0.

[20] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, JillÂăP. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417 – 425, 2015.

[21] Maria Tiana, Barbara Acosta-Iborra, Laura Puente-Santamaría, Pablo Hernansanz-Agustin, Rebecca Worsley-Hunt, Norma Masson, Francisco García-Rio, David Mole, Peter Ratcliffe, Wyeth W Wasserman, Benilde Jimenez, and Luis delÂăPeso. The sin3a histone deacetylase complex is required for a complete transcriptional response to hypoxia. *Nucleic Acids Research*, 46(1):120–133, 2018.

[22] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, Dec 2014.

[23] Andrew T. Kwon, David J. Arenillas, Rebecca Worsley Hunt, and Wyeth W. Wasserman. oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-Representation Across Genes or ChIP-Seq Datasets. *G3: Genes, Genomes, Genetics*, 2(9):987–1002, 2012.

[24] Zhenjia Wang, Mete Civelek, Clint L Miller, Nathan C Sheffield, Michael J Guertin, and Chongzhi Zang. Bart: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics*, 34(16):2867–2869, 2018.

[25] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin vanÂăderÂăLee, Adrien Bessy, Jeanne Chèneby, Shubhada R Kulkarni, Ge Tan, Damir Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266, 11 2017.

[26] Su Wang, Chongzhi Zang, Tengfei Xiao, Jingyu Fan, Shenglin Mei, Qian Qin, Qiu Wu, Xujuan Li, Kexin Xu, Housheng Hansen He, Myles Brown, Clifford A. Meyer, and X. Shirley Liu. Modeling cis-regulation with a compendium of genome-wide histone h3k27ac profiles. *Genome Research*, 26(10):1417–1429, 2016.

[27] Alexander Lachmann, Andrew D. Rouillard, Caroline D. Monteiro, Gregory W. Gundersen, Kathleen M. Jagodnik, Matthew R. Jones, Maxim V. Kuleshov, Michael G. McDermott, Nicolas F. Fernandez, Qiaonan Duan, Sherry L. Jenkins, Simon Koplev, Zichen Wang, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016.

[28] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291, 09 2011.

[29] Andreas Keller, Christina Backes, and Hans-Peter Lenhof. Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics*, 8(1):290, Aug 2007.

# A

# Manual de utilización

TFEA.ChIP is designed to take the output of a differential expression analysis and identify TFBS enriched in the list of DE genes. In the case of the analysis of association, the only required input is a set of DE genes and, optionally, a set of control genes whose expression is not altered by the experimental conditions under study. For the GSEA analysis a ranked list of genes is required. This is supplied as a dataframe containing a column with gene names and a numerical column with the ranking metric, which typically are log-fold change or p-values for the gene expression changes in the two conditions under evaluation. For illustration purposes we will derive the input required for both analysis from a table containing the following field columns:

- Gene name (*Genes*). Internally the package uses Entrez IDs, but translating from Gene Symbols and ENSEMBL IDs is available.

- Log2 Fold Change (*Log2FoldChange*), indicating the difference in expression for each gene in the two experimental conditions being compared.

- p-value (*pvalue*) or adjusted p-value (*pval.adj*) for the difference in gene expression between the two conditions.

The output of popular packages, such as DESeq2, for detection of differentially expressed genes from the analysis of count data from RNA-seq experiments produce tables with this information. The *hypoxia_ DESeq* and *hypoxia* datasets are the output of a differential expression analysis performed on an RNAseq experiment analyzing the response to hypoxia of endothelial cells[21] deposited at the NCBI's GEO repository (GSE89831).

To extract the information from a DESeqResults object or a data frame the function *preprocessInputData* is available. Using the option *from.Mouse = TRUE* will translate mouse gene IDs to their equivalent human gene ID:

```
library(TFEA.ChIP)
data( "hypoxia_DESeq", "hypoxia", package="TFEA.ChIP" ) # load example datasets
hypoxia_table <- preprocessInputData( hypoxia_DESeq )
```

After running *preprocessInputData*, your dataset will be ready to use with the rest of the package; gene names will be in Entrez Gene ID format and the resulting table is sorted by log2(Fold Change).

## A.1. Analysis of the association of TFBS and differential expression

### A.1.1. Identification of DE genes

As indicated before, for this analysis, we must provide a list of genes are considered differentially induced and a list of control genes whose expression is not altered in the analyzed experiment. For that we will use the function *Select_genes*:

```
    #extract vector with names of upregulated genes
Genes.Upreg <- Select_genes( hypoxia_table, min_LFC = 1 )
    #extract vector with names of non-responsive genes
Genes.Control <- Select_genes( hypoxia_table,
    min_pval = 0.5, max_pval = 1,
    min_LFC = -0.25, max_LFC = 0.25 )
```

### A.1.2. Translate the gene IDs to Entrez Gene IDs

In case the input dataset cannot be preprocessed or the user is interested in analyzing a particular set of genes that doesn't come from the input dataset, translating the IDs to Entrez Gene ID format is required. To that end its available the function *GeneID2entrez*:

```
#Conversion of hgnc to ENTREZ IDs
GeneID2entrez( gene.IDs = c("EGLN3","NFYA","ALS2","MYC","ARNT" ) )
# To translate from mouse IDs:
# GeneID2entrez( gene.IDs = c( "Hmmr", "Tlx3", "Cpeb4" ), from.Mouse = TRUE )
```

### A.1.3. Association analysis

In this step, we will construct a contingency table for each of the factors stored in the internal database categorizing the DE (DE_yes) and background (DE_no) genes according to the presence or absence of binding sites:

| | TFbound_yes | TFbound_no |
|---|---|---|
| DE_yes | number y/y | number y/n |
| DE_no | number n/y | number n/n |

Then, we will apply Fisher's exact test to each contingency table to test the null hypothesis that factor binding and differential expression are independent. In addition, to the raw p-values the function also return the FDR-adjusted values to correct for multiple testing.

```
CM_list_UP <- contingency_matrix( Genes.Upreg, Genes.Control )
    #generates list of contingency tables, one per dataset
result_UP <- getCMstats( CM_list_UP )
    #generates list of p-values and OR from association test
```

In this example, all 1122 datasets in the internal database were used in the analysis. However, we can restrict the analysis to a specific subset of the database and/or a given set of transcription factors. To this end we can produce and index of the tables of interest with the function *get_chip_index* and pass this index as an additional argument to *contingency_matrix*. Finally, note that the list of control genes is optional. If not supplied, all human genes not present in the test list will be used as control. Thus, we could restrict the analysis to the datasets generated by the ENCODE project and use all non-DE genes as control:

```
    #restrict the analysis to datasets assaying these factors:
chip_index <- get_chip_index( TFfilter = c( "HIF1A","EPAS1","ARNT" ) )
    # Or select ENCODE datasets only:
chip_index <- get_chip_index( encodeFilter = TRUE )
```

To know more about the experiments included in TFEA.ChIP's database or the conditions of a particular experiment, load the metadata table included using *data("MetaData", package = "TFEA.ChIP").*

To summarize the results by transcription factor use *rankTFs*. This function performs Wilcoxon rank-sum test or GSEA to test whether ChIPs belonging to the same TF are, as a group, significantly enriched / depleted in the results of the analysis. Be aware that in the case of transcription factors whose behavior is dependent on cellular context, integrating the results of all the related ChIPs might conceal its enrichment in a particular set of experimental conditions.

```
TF_ranking <- rankTFs( result_UP, rankMethod = "gsea", makePlot = T )
```

### A.1.4. Plot results

The table of results generated by *getCMstats* can be parsed to select candidate TF. The function *plot_CM uses* the package *plotly* to generate an interactive plot representing the p-value against the odd-ratio that is very helpful to explore the results.

```
plot_CM( result_UP ) #plot p-values against ORs
```

In fact, the exploration of this graph shows a strong enrichment for several HIF datasets, as expected for the *hypoxia dataset*. This can be clearly shown by highlighting the datasets of interest:

```
HIFs <- c( EPAS1="EPAS1",HIF1A="HIF1A",ARNT="ARNT" )
col <- c( "red","blue","green" )
plot_CM( result_UP, specialTF = HIFs, TF_colors = col )
    #plot p-values against ORs highlighting indicated TFs
```

## A.2. Gene Set Enrichment Analysis

### A.2.1. Generate a sorted list of ENTREZ IDs

The GSEA analysis implemented in the TFEA.ChIP package requires as input a sorted list of genes. By default, the function *preprocessInputData* will sort genes according to log fold change in descending order. However, they could be sorted by any numerical parameter including p-value. If you want to generate your custom gene list with other parameters, remember to make sure the gene IDs are in Entrez Gene ID format or translate them with *GeneID2Entrez*.

### A.2.2. Select the ChIP-Seq datasets to analyze

By default, the analysis will include all the ChIP-Seq experiments available in the database. However, this analysis might take several minutes to run. To restrict the analysis to a subset of the database we can generate an index variable and pass it to the function *GSEA_run*. This will limit the analysis to the ChIP-Seq datasets of the user's choosing. This index variable can be generated using the function *get_chip_index* and allows the user to select the whole database,

the set of ChIP-Seq experiments produced by the ENCODE project (."encode") or a specific subset of transcription factors (as a vector containing the TF names).

```
chip_index <- get_chip_index( TFfilter = c( "HIF1A","EPAS1","ARNT" ) )
    #restrict the analysis to datasets assaying these factors
```

### A.2.3. Run the GSEA analysis

The function *GSEA_run* will perform a GSEA-based analysis on the input gene list. This function is based on the R-GSEA R script bundle written by the GSEA team at the Broad Institute of MIT and Harvard[17]. The output of the analysis depends on the variable *get.RES*:

- When **False**, the function returns a data frame storing maximum Enrichment Score and associated p-value determined for each dataset included in the analysis.

- When **True**, the function returns a list of three elements. The first element (*Enrichment.table*) is the enrichment data frame previously mentioned. The second element (*RES*) is a list of vectors containing the *Running Enrichment Score* values (see GSEA documentation) for each of the sorted genes tested against each one of the analyzed ChIP datasets. The third element (*indicators*) is a list of vectors indicating if the sorted genes were bound by the factor analyzed in each ChIP dataset.

```
GSEA.result <- GSEA_run( #run GSEA analysis
    hypoxia_table$Genes, hypoxia_table$log2FoldChange,
    chip_index, get.RES = TRUE)
```

The list of results can be restricted to a given set of transcription factors by setting the variable *RES.filter*.

### A.2.4. Plotting the results

TFEA.ChIP includes two functions that use the package *plotly* to generate interactive html plots of your GSEA results: *plot_ES* and *plot_RES*.

- Plot Enrichment Scores with *plot_ES*:

We can choose to highlight ChIP-Seq from specific transcription factors plotting them in a particular color.

```
TF.hightlight <- c( EPAS1="EPAS1", ARNT="ARNT" )
col <- c( "red","blue" )
plot_ES( GSEA.result,
    LFC = hypoxia_table$log2FoldChange,
    specialTF = TF.hightlight, TF_colors = col)
```

- Plot Runing Enrichment Scores with *plot_RES*:

This function will plot **all** the RES stored in the *GSEA_run* output. It is only recommended to restrict output to specific TF and/or datasets by setting the parameters *TF* and/or *Accession* respectively:

```
plot_RES( GSEA.result,
    LFC = hypoxia_table$log2FoldChange, TF = c( "EPAS1" ),
    Accession = c( "GSM2390642", "GSM1642766" ) )
```

## A.3. Building a TF-gene binding database

If the user wants to generate their own database of ChIPseq datasets, the functions *txt2gr* and *GR2tfbs_ db* automate most of the process. The required inputs are:

- A Metadata table (storing at least, Accession ID, name of the file, and TF tested in the ChIP-Seq experiment). The metadata table included with this package has the following fields: Name, Accession, Cell, Cell Type, Treatment, Antibody, and TF.

- A folder containing ChIP-Seq peak data, either in «.narrowpeak» format or the MACS output files «_peaks.bed» (a format that stores chr, start, end, name, and Q-value of every peak).

### A.3.1. Filter peaks from source and store them as a GRanges object

Specify the folder where the ChIP-Seq files are stored, create an array with the names of the ChIP-Seq files, and choose a format. Set a *for* loop to convert all your files to GenomicRanges objects using *txt2GR*. Please note that, by default, only peaks with an associated p-value of 0.05 (for narrow peaks files) or 1e-5 (for MACS files) will be kept. The user can modify the default values by setting the alpha argument to the desired threshold p-value.

```
folder <- "~/peak.files.folder"
File.list<-dir( folder )
format <- "macs"

gr.list <- lapply(
    seq_along( File.list ),
    function( File.list, myMetaData, format ){
        tmp<-read.table( File.list[i], ..., stringsAsFactors = FALSE )
        file.metadata <- myMetaData[ myMetaData$Name == File.list[i], ]
        ChIP.dataset.gr<-txt2GR(tmp, format, file.metadata)
    return(ChIP.dataset.gr)
    },
    File.list = File.list,
    myMetadata = myMetadata,
    format = format
)
    # As an example:
data( "ARNT.peaks.bed","ARNT.metadata", package = "TFEA.ChIP" )
ARNT.gr <- txt2GR( ARNT.peaks.bed, "macs1.4", ARNT.metadata )
```

### A.3.2. Assign TFBS peaks from ChIP dataset to specific genes

The function *GR2tfbs_ db* assigns the TFBS peaks in the ChIP datasets stored in *gr.list* to a gene. To this end, a ChIP peak overlaping a Dnase HS region receive the gene label associated to that region. By default the function also assigns the gene name when the ChIP peak does not overlap a Dnase region but maps at less than 10 nucleotides from it. This behaviour can be modified by setting the argument *distanceMargin* to the desired value (by default distanceMargin = 10 bases).

The function, accepts any Genomic Range object that includes a metacolumn with a gene ID (stored in the `@elementMetadata@listdata[["gene_id"]]` slot of the object) for each genomic segment. For example, asignation of peaks to genes can be done by providing a list of all the genes in the genome:

```
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
data( "gr.list", package="TFEA.ChIP") # Loading example datasets for this function
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
Genes <- genes( txdb )
TF.gene.binding.db <- GR2tfbs_db( Genes, gr.list, distanceMargin = 0 )
```

### A.3.3. Generation of the TFBS database

The function *makeTFBSmatrix* generates a binary matrix with a row for each human gene and a column for each independent ChIPseq dataset. The cell of the matrix contain a value of 1 if the gene is bound by the TF and 0 otherwise. This matrix and the metadata table can be used instead of the default TFBS database.

```
data( "tfbs.database", package = "TFEA.ChIP" ) # Loading example datasets for this
    function
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
gen.list <- genes( txdb )$gene_id # selecting all the genes in knownGene

myTFBSmatrix <- makeTFBSmatrix( gen.list, tfbs.database )
myTFBSmatrix[ 2530:2533, 1:3 ] # The gene HMGN4 (Entrez ID 10473) has TFBS for this
    three ChIP-Seq datasets
```

### A.3.4. Sustitute the default database by a custom generated table

At the beginning of a session, use the function *set_user_data* to use your TFBS binary matrix and metadata table with the rest of the package.

```
set_user_data( binary_matrix = myTFBSmatrix, metadata = myMetaData )
```